DOCUMENT RESUME

ED 464 951 TM 033 882

AUTHOR Good, Robert
TITLE Using Discriminant Analysis as a Method of Combining Multiple Measures of Student Performance.
PUB DATE 2002-04-00
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Attendance; Grade Point Average; Mathematics Tests; *Middle School Students; Middle Schools; Reading Tests; Standardized Tests
IDENTIFIERS Colorado Student Assessment Program; *Multiple Measures Approach; *Predictive Discriminant Analysis

ABSTRACT

Current methods of combining multiple measures of student performance have been subjective and have lacked evidence of external validity. This study examined the use of predictive discriminant analysis (PDA) as a means of finding the best combination of variables to predict performance on a standards-based mathematics assessment. Data from 261 eighth-grade middle school students were analyzed. Predictor variables included: (1) attendance rate; (2) grade point average; (3) a locally developed mathematics assessment; (4) a standardized reading assessment; and (5) the mathematics and reading subtests of a standardized achievement test. The grouping variable was the attained level on the Colorado Student Assessment Program (CSAP). Results indicate that the four performance levels posited by the CSAP authors were not predicted accurately. After the data were dichotomized into Proficient and Not Proficient categories, the combination of grade point average, locally developed mathematics assessment, standardized reading assessment, and two of the mathematics achievement subtests was the best in terms of accounted variance and classification accuracy. The use of a PDA for the purposes of objectively combining multiple measures of student performance was supported. The information generated by this method and its advantages are discussed in the context of educational decision making. (Contains 36 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 464 951

Using Discriminant Analysis as a Method of Combining Multiple Measures of Student Performance

Robert Good

Durango, CO

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

R. Good

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the Annual Meeting of the

American Educational Research Association

New Orleans, LA. April, 2002.

TM033882

Note. This paper is based in part on the author's doctoral dissertation - a complete citation is given in the reference section.

BEST COPY AVAILABLE

Abstract

Using a variety of indicators as the basis of important educational decisions is clearly preferable to relying on a single observation or assessment. However current methods of combining multiple measures of student performance have been subjective and have lacked evidence of external validity. This study examined the use of predictive discriminant analysis (PDA) as a means of finding the best combination of variables to predict performance on a standards-based mathematics assessment.

Data from 261 8th -grade middle school students were analyzed. Ten predictor variables included attendance rate, grade point average, a locally developed mathematics assessment, a standardized reading assessment, and the mathematics and reading subtests of a standardized achievement test. The grouping variable was the attained level on the Colorado Student Assessment Program (CSAP).

Results indicated that the 4 performance levels posited by the CSAP authors were not predicted accurately. After dichotomizing the data into Proficient and Not Proficient categories, the combination of grade point average, locally developed mathematics assessment, standardized reading assessment, and 2 of the mathematics achievement subtests was the best in terms of accounted variance and classification accuracy. The use of a PDA for the purposes of objectively combining multiple measures of student performance was supported. The information generated by this method and its advantages are discussed in the context of educational decision making.

Introduction

One of the cornerstones of public education has been the evaluation of student classroom performance utilizing a variety of methods. Classroom activities, summative tests, periodic quizzes, and teacher observations are just some of the measures teachers use throughout the year to assess student learning and performance. These measures are then typically combined in some manner to arrive at a summative conclusion or grade that is based on the many assessments and observations that have taken place over the course of the instructional period. Although the potential consequences attached to the conclusion or grade itself may be significant, the relative magnitude of a single assessment or observation is diminished in light of the entire body of evidence used to reach the conclusion.

However in response to calls for greater accountability of public schools, there has been an increased interest in isolated, external measures of student performance. Additionally, the recent widespread use of educational standards has cast more attention on external assessment systems designed to measure student proficiency in specific content areas. Significant rewards and consequences have been attached to the outcomes of many of these systems (Education Week, 1999), thereby making them high-stakes in nature (Lewis, 2000).

When making important educational decisions based on student performance, both the quantity and quality of the information used have been important factors in the confidence that can be placed in the accuracy of the decisions. As Mehrens (1990) stated, "In general, the more data that are gathered, the better the decision is likely to be. Certainly it is conventional psychometric wisdom that one should use

more than one piece of data as the basis for important decisions" (p. 322). Further support for the use of multiple measures has been stated explicitly in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999), wherein Standard 13.7 states:

> In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision. (p. 146)

Yet while the principle of combining multiple measures of student performance as the basis for making decisions has been used by classroom teachers for years, its use within the context of educational reform has been, at best, limited.

Methods of Combining Multiple Measures

Mehrens (1990) distinguished two basic approaches of combining multiple measures of performance. The clinical approach involves a subjective process of simply looking at the data collectively and then making judgments based on a general impression. The statistical approach applies a series of weights to each of the components and then combines them mathematically. Of the specific methods used in this approach, only a few yield direct evidence of external validity by assessing the accuracy of the equation (Ryan & Hess, 1999). Of the two approaches, Mehrens (1990) stated that the statistical approach yields results that are more accurate than the clinical approach. However, current literature indicates that subjective methods are more prevalent than objective methods that also generate evidence of external validity.

Ryan and Hess (1999) described several specific methods of combining multiple measures using established minimum performance levels as the basis for decisions. Disjunctive procedures assign a passing score to those who meet the minimum requirements on any of the measures used. Conversely, conjunctive procedures require that individuals meet the minimum requirements on all the measures used. The compensatory approach blends disjunctive and conjunctive procedures by allowing high performance on some of the measures to make up for low performance on others. This approach is particularly appropriate in high-stakes environments where the policy goal of high achievement must be balanced against the political realities of not attaining such achievement (Novak, Winters, & Flores, 2000).

Other methods of combining multiple measures utilize the raw scores or distributional characteristics of each of the component measures in absence of an external criterion. Methods that include using raw scores, test length and difficulty, and variance equalization (Wang & Stanley, 1970), inverses of standard error (Gulliksen, 1950), and test reliability (Rudner, 2000) as the basis for assigning component weights have been discussed. However, there is general consensus in the literature that the methods used to combine multiple measures in absence of an external criterion are inferior to those that

utilize an appropriate external measure. Gulliksen (1950) expressed this sentiment by stating, "*If a criterion is available, multiple correlation methods give the best weights for predicting that criterion*" (p. 330, italics in original). Furthermore, Wainer and Thissen (1993) contended that an external criterion should be used when combining items from different test formats (i.e., multiple choice and constructed response). This issue becomes particularly relevant in light of the current interest in using performance assessments for educational accountability purposes.

While the generalized approach utilizing an external criterion is based on the multiple correlation (e.g., multiple regression), specific methods have been recently applied to this problem. Ryan and Hess (1999) described two such methods. The first utilized various geometric distances (viz., raw, standardized, or Mahalanobis) from a set of criterion variables to predict group membership and assess classification accuracy. The second method utilized a discriminant function to yield similar information as the measures of geometric distance, but also included a measure of external validity.

Current Applications of Combining Multiple Measures

The accuracy of assessment systems that combine relatively few measures of performance is very dependent upon the measures themselves as well as the method used to combine them. The current literature indicates much more concern with the component measures themselves than with the method of combining them. Jang (1998) described a school district's effort to comply with a statutory requirement of using multiple measures for a standards-based accountability system related to Title I funds. Although the district used several standardized and performance-based assessments, the method used to weight the components was based simply on the proportion of item type within the overall battery.

Hohn and Veitch (1999) described a method of combining data for the purpose of measuring basic literacy levels in compliance with a state mandate that involved the blending of conjunctive and disjunctive procedures. This method stipulated that satisfactory performance be defined as proficient performance on two out of three measures. This criterion is questionable, however, in that even though the measures used had psychometric support, there was no equating done between the measures. The question of whether or not each pair of possible combinations is comparable to the other combinations is an important one.

Purpose of Study

A review of current applications of multiple measures suggested that a variety of subjective, policy-based methods have been used to combine assessment components. While methods do exist for deriving component weights statistically (e.g., Ryan & Hess, 1999), there was little evidence that these methods have been used in actual assessment systems even though these methods are preferable both in terms of predictive accuracy (Gulliksen, 1950) and as applied to assessments employing more than one format (Wainer & Thissen, 1993). There was also support for the contention that statistical methods

provide a better opportunity to explore a combination of multiple measures more thoughtfully and better supported the inferences and conclusions drawn in light of the accuracy of the combination. The problem addressed in this study was that the methods currently used for combining multiple measures of student performance were often subjective and lacked evidence of external validity. Therefore, the purpose of this study was to use a discriminant analysis to combine a set of multiple measures in a manner that best predicted scores on a statewide, standards-based mathematics assessment. Additionally, this study sought to explore the types of information generated by this method.

Method

Subjects

The subjects for this study were 261 8th-grade students from two middle schools in southwest Colorado, fictitiously named Edison Middle School and Madison Middle School. Both schools served students in grades six through eight. The students were on teams consisting of approximately 75-100 students with three or four teachers assigned to the core areas of reading, language arts, mathematics, science, and social studies. The students were enrolled in one of four math classes: Remedial, Pre-Algebra, Algebra, or Geometry. Table 1 presents a crosstabulation of school by math course.

Table 1

School by Math Course Crosstabulation

	Math Course				
School	Remedial	Pre-Algebra	Algebra	Geometry	Total
Madison	1	81	52	2	136
Edison	12	67	40	5	124
Total	13	148	92	7	260

Note. Does not include one missing case from Madison.

This study utilized data from the 1999-2000 school year. The total number of students in the study was 261, including 146 males and 115 females. The students ranged in age from 160 mos. to 193 mos. with a mean of 170.3 (SD = 5.1). The students were representative of a wide range of economic backgrounds, however ethnic minorities represented only 12.3% of the total sample.

Data Analysis

The statistical method employed in this study was multiple discriminant analysis (MDA). It is a complex, yet versatile, multivariate procedure that allows the researcher the flexibility to examine data from different perspectives before drawing conclusions or making inferences. MDA is appropriate when group membership is either predicted from - or interpreted by - a set of predictor variables (Klecka,

1980). In this study, MDA was employed to investigate how well a set of independent (predictor) variables predicted group membership on a standards-based assessment.

Several authors (Buras, 1996; Huberty, 1984, 1994; Huberty and Barton, 1989; Woldbeck, 1998) have distinguished between two types of discriminant analysis. Descriptive discriminant analysis (DDA) seeks "... to study and explain group separation or group differences" (Buras, 1996, p. 6). In this case, the grouping variable is independent and the predictor variables are dependent, thus indicating a MANOVA design (Huberty, 1994). The linear discriminant functions (LDFs) are relevant here because they indicate the relative weight of each predictor variable in a way that best separates the groups (Woldbeck, 1998). In DDA group separation, not classification, is of interest so that the differences between the groups can be explained.

When the purpose is to classify subjects, predictive discriminant analysis (PDA) is the appropriate type. PDA seeks to use a set of predictor (independent) variables that maximizes the classification accuracy on the grouping (dependent) variable, regardless of the weights given by the LDFs (Huberty, 1984). In the case of PDA, the linear classification functions (LCFs) are used to determine group membership (Woldbeck, 1998). Because this study investigated the set of variables that best predicts performance on the CSAP, a PDA was used.

Variable Selection

There are two common methods of variable selection in MDA: stepwise and all-possible subsets. Stepwise procedures utilize a set of criteria for variable inclusion in the functions. Only the variables that meet the criteria at each step are used in the final result. The all-possible subsets method examines all combinations of the variables. For example, in the case of a variable set consisting of A, B, and C, the combinations of A, B, C, A-B, A-C, B-C, and A-B-C would all be examined.

Although much less cumbersome, there have been some serious concerns raised over the use of stepwise procedures (see Thompson, 1995; Whitaker, 1997). Because these concerns were serious enough to offset the convenience and ease of a stepwise procedure, this study considered all-possible subsets of predictor variables from which the final variable set was selected.

Selection Criteria and Cross-Validation Methods

The criteria used to make the predictor variable selection included hypothesis testing and classification accuracy. For the hypothesis testing, the Wilks lambda criterion (Λ) is used to test the significance of the difference between the group means (Lindeman, Merenda, & Gold, 1980). Klecka (1980) stated that because Λ is distributionally similar to chi-square, the proportion of cases farther from the group centroid has interpretive value as a measure of similarity to other cases in the assigned group. It

is also important to note that Λ and the proportion of accounted variance (R^2) always add to one ($\Lambda + R^2 = 1$), therefore small values of Λ imply greater explanatory power of the discriminant functions (Pedhazur, 1982).

Initial classification accuracy (hit rate) was computed through an internal analysis by simultaneously deriving the functions and classifying the data. This initial analysis tends to produce overly optimistic hit rates (Hirst, 1996; Huberty, 1994). A more conservative cross-validation procedure is the leave-one-out (L-O-O) method described by Huberty (1994). This method involves the repeated process of deleting a single case, and then classifying that case using the LCFs from the remaining data.

Error rate estimation can also be accomplished using the obtained posterior probabilities of group membership. Dillon and Goldstein (1984) pointed out that an "overall error rate estimator" of a multiple discriminant analysis can be "… calculated from the average of the maximum posterior probabilities [M-P-P] for each observation" (p. 408). Further, the authors noted that since the actual classification of a subject is not used in calculating the posterior probabilities, this error rate estimator is useful for the interpretation of new cases. Therefore, the final combination of predictor variables was chosen after considering the statistical significance of the canonical function, the observed initial hit rate, the L-O-O cross-validation analysis, and the overall error rate using the average posterior probability.

Instrumentation

Grouping variable. The Mathematics portion of the Colorado Student Assessment Program (CSAP) (Colorado Department of Education [CDE], 2000) was used as the grouping variable. It consisted of both selected response (56% of items) and constructed response (44% of items). Scaled scores were generated and then converted into one of four performance levels (Advanced, Proficient, Partially Proficient, Unsatisfactory).

Predictor variables. The initial set of predictor variables included several subtests on the Iowa Test of Basic Skills (ITBS), Level 14, Form K (Hoover, Hieronymus, Frisbie, & Dunbar, 1993), the Gates-MacGinitie Reading Test (GMRT), 4th Edition (MacGinitie, MacGinitie, Maria, & Dreyer, 2000), a locally developed mathematics assessment, attendance data, and student grades. For purposes of notational clarity, predictor variables are referenced as follows: Attendance Rate (ATT), Mathematics Grade Point Average (GPA), District Mathematics Assessment (DMA), ITBS Concepts and Estimation (CE), ITBS Problem Solving (PS), ITBS Computation (COMP), ITBS Total Math (MATH), ITBS Reading (RDG), ITBS Total Core Battery (CORE), and GMRT Extended Scale Score (GMRT). A summary of the psychometric properties of the grouping and predictor variables, along with specific variable scoring and coding details are presented in Good (2001).

Data Processing and Analysis

The data for this study were analyzed using the Statistical Package for the Social Sciences (SPSS), version 8.0 (SPSS, 1998a). Preliminary analyses included a verification of the multivariate assumptions of normality, linearity, and minimal multicollinearity. As suggested by Huberty (1994), multiple one-way ANOVAs of each predictor variable were conducted to investigate univariate differences across the levels of the CSAP. While a non-significant ($F < 1$) result did not necessarily exclude a variable from further consideration, the information obtained helped identify the final set of predictor variables.

The predictor variables were entered into a PDA in an all-possible subsets manner. Each analysis utilized actual group membership to establish prior probabilities. Accounted variance estimates, initial classification accuracy estimates, and cross-validation accuracy estimates using an L-O-O procedure were generated. Maximum posterior probabilities were averaged to produce an error rate estimator (M-P-P). The accounted variance, the classification accuracy estimates, and the M-P-P estimates were used to select the final set of predictor variables.

Results

The sample distribution across the four levels of the CSAP was as follows: Advanced $n = 29$ (11%); Proficient $n = 80$ (31%); Partially Proficient $n = 104$ (40%); Unsatisfactory $n = 47$ (18%). Table 2 gives a summary of student performance on the predictor variables by CSAP level.

The relatively low N of cases (116) associated with GMRT was due to the fact that Edison reported scores in grade level equivalents rather than extended scale scores (ESSs). Grade level equivalents were not usable because of the clustering of scores at the upper end of the distribution that did not occur with the associated ESSs.

Preliminary PDA Analyses

Results showed that ATT violated the distributional assumptions (i.e., normality and equality of predictor covariance matrices across the grouping variable) and was therefore dropped from consideration. Taking all combinations of the nine remaining predictor variables resulted in a total of 511 possible combinations. Eliminating those that included pairs of variables that correlated higher than .8 in order to limit construct redundancy reduced the number of combinations to 180. Each combination was entered directly into a PDA. For purposes of initial variable subset reduction, measures of accounted variance [(1- A) X 100%], initial classification accuracy, and L-O-O cross-validation accuracy were examined (see Table 3).

Table 2

Mean Student Performance on Predictor Variables by CSAP Level

Predictor Variable	*N*	Min/Max	*M*(*SD*)	CSAP Level: Advanced *n*	Advanced *M*(*SD*)	Proficient *n*	Proficient *M*(*SD*)	Partially Proficient *n*	Partially Proficient *M*(*SD*)	Unsatisfactory *n*	Unsatisfactory *M*(*SD*)
ATT	260	69.4/100	94.9 (4.8)	29	94.5 (4.6)	80	95.8 (3.7)	103	94.7 (4.8)	47	94.0 (6.3)
DMA	231	2/40	22.7 (9.1)	26	33.2 (4.4)	73	27.2 (6.3)	95	20.0 (7.4)	37	13.1 (7.8)
GMRT	116	486/630	569.7 (29.0)	12	596.0 (16.3)	35	586.0 (23.4)	46	561.6 (24.6)	22	545.8 (26.6)
GPA	260	0.17/4.00	2.98 (.89)	29	3.66 (0.50)	80	3.44 (0.54)	104	2.74 (0.85)	46	2.32 (1.01)
COMP	257	163/333	256.6 (32.9)	29	290.2 (26.5)	80	272.5 (22.7)	104	247.2 (26.7)	47	229.1 (32.7)
CE	258	161/337	270.0 (32.6)	29	307.7 (17.6)	80	289.5 (17.6)	104	263.9 (19.9)	47	227.3 (29.4)
PS	258	157/365	277.3 (36.8)	29	310.7 (27.7)	80	296.3 (19.8)	104	272.0 (26.0)	47	236.3 (42.7)
MATH	258	166/334	273.5 (32.9)	29	309.6 (21.1)	80	292.8 (15.8)	104	268.0 (20.7)	47	230.9 (31.8)
RDG	258	168/331	271.9 (30.4)	29	294.1 (16.4)	80	283.2 (20.1)	104	267.3 (20.4)	47	238.4 (31.9)
CORE	258	166/331	271.9 (30.9)	29	305.4 (13.5)	80	288.0 (18.5)	104	267.5 (18.9)	47	233.8 (31.6)

Table 3

Summary of Initial PDA Results For All Subset Combinations

Variables in Combination	Number of Combinations	Analysis Indicator: Accounted Variance *M*	Accounted Variance *SD*	Initial Accuracy *M*	Initial Accuracy *SD*	L-O-O Accuracy *M*	L-O-O Accuracy *SD*
1	9	35.8%	12.7	46.7%	10.1	44.4%	13.0
2	30	47.6%	9.2	51.7%	7.3	48.7%	7.3
3	54	53.9%	7.6	54.3%	5.6	48.5%	6.1
4	49	58.6%	5.6	55.9%	4.8	49.1%	5.7
5	28	62.8%	3.7	57.9%	3.9	49.7%	4.8
6 & 7	10	66.0%	1.8	60.8%	3.0	52.0%	3.4
Total	180	55.3%	9.9	54.9%	6.4	48.9%	6.6

Note. Accounted variance = [(1 - Λ) X 100%]; Initial Accuracy = % of cases correctly classified internally; L-O-O = % of cases correctly classified by the Leave-One-Out cross-validation procedure.

The poor accounted variance and low classification accuracy estimates of the four-group solutions were of great concern. Errors associated with decisions based on these solutions would have been unacceptably high. Given that the Unsatisfactory and Advanced levels included only 29% of the cases, the CSAP performance levels were recoded dichotomously so that those in the Advanced and Proficient levels were assigned to a Proficient level, and those in the Partially Proficient and Unsatisfactory levels were assigned to a Not Proficient level. The distribution across the recoded levels of the CSAP was as follows: Proficient n = 109 (42%); Not Proficient n = 151 (58%).

Table 4 presents a summary of variance and accuracy estimates for the 180 combinations of predictor variables analyzed using the two-group CSAP scores. Average accounted variance and accuracy estimates supported the decision to dichotomize the grouping variable. These indicators improved over the four-group solution and were high enough to be useful for decision-making purposes in educational settings. Additionally, they supported the contention that multiple measures were better than any single measure. While the benefit of multiple measures would reasonably level off at some point, the improvement over a single measure was dramatic.

Table 4

Summary of PDA Results for All Subset Combinations Using Two-Group CSAP Scores

Variables in Combination	Number of Combinations	Analysis Indicator					
		Accounted Variance		Initial Accuracy		L-O-O Accuracy	
		M	*SD*	*M*	*SD*	*M*	*SD*
1	9	30.9%	6.4	75.4%	4.8	75.2%	5.0
2	30	37.3%	7.3	79.3%	4.1	78.4%	8.7
3	54	41.9%	5.9	80.8%	3.3	79.3%	3.8
4	49	44.9%	4.7	82.0%	3.1	79.7%	3.5
5	28	47.7%	3.3	83.0%	2.6	80.3%	3.2
6 & 7	10	50.0%	1.8	84.6%	2.4	80.8%	2.1
Total	180	42.8%	7.0	81.2%	3.9	79.3%	3.9

Note. Accounted variance = [(1 - A) X 100%]; Initial Accuracy = % of cases correctly classified internally; L-O-O = % of cases correctly classified by the Leave-One-Out cross-validation procedure.

Selected Combinations

Initial analyses of the 180 combinations included revealed many that were comparable in terms of accounted variance and classification accuracy. Therefore, the decision as to which combination would be used was based on somewhat subtle differences in variance and accuracy. The inclusion of DMA and

GMRT affected the analyses considerably due to the number of missing cases within these variables. Consequently, the final set of combinations selected considered the number of valid cases that resulted from the inclusion of DMA and GMRT, as well as the accounted variance and classification accuracy. Three combinations were selected for further analyses based on the strength of their estimates:

Analysis One: GPA-COMP-CE-PS-RDG

Analysis Two: GPA-DMA-COMP-CE-RDG

Analysis Three: GPA-DMA-CE-PS-GMRT.

Procedural SPSS syntax for each analysis called for a direct, linear PDA. Prior probabilities were based on group proportions given by the known distribution of actual scores. Only those cases with valid values for each predictor variable in a given analysis were included. The pooled covariance matrices were used, and the data were cross-validated using an L-O-O procedure. Table 5 presents a summary of the PDA for each of these combinations.

Table 5

Summary of Discriminant Analysis for Selected Combinations

Combination	*N*	Accounted Variance	Initial Classification Accuracy	L-O-O	M-P-P
GPA-COMP-CE-PS-RDG	255	45.9%*	86.7%	85.9%	81.6%
GPA-DMA-COMP-CE-RDG	226	50.1%*	86.7%	85.8%	83.8%
GPA-DMA-CE-PS-GMRT	98	50.3%*	86.7%	85.7%	83.2%

Note * - Associated Wilks' Λ sig. ($p < .01$). Initial Accuracy = % of cases correctly classified internally; L-O-O = % of cases correctly classified by the Leave-One-Out cross-validation procedure; M-P-P = mean posterior probability of group membership.

In general, the three combinations were very comparable. The variance accounted by Analysis One was slightly less than that of Analyses Two and Three, however accuracy and probability estimates of all three analyses were consistent and acceptable. The mean L-O-O estimate was 85.8% and the mean M-P-P estimate was 82.9%.

Relative component strength of each predictor variable could be assessed through the discriminant loadings and the LDF coefficients (see Table 6). The standardized LDF coefficients were used in this context as they indicated the relative importance of each predictor while controlling for differences in measurement units among the predictor variables. The discriminant loadings ranged from moderate (.415) to strong (.887) in magnitude supporting the notion that the predictor variables related as a variate to the construct measured. Additionally, the standardized LDF coefficients indicated that a student's score on CE impacted group placement considerably. For all three analyses, changes in CE performance (Z_{CE} = .647, .546, & .657 for Analyses One, Two, & Three, respectively) would impact a

student's position relative to the group centroids (classification group mean) more than any other predictor variable.

Table 6

Discriminant Loadings and Standardized LDF Coefficients

Analysis 1			Analysis 2			Analysis 3		
Variable	LDF	Loading	Variable	LDF	Loading	Variable	LDF	Loading
GPA	.413	.887	GPA	.348	.814	GPA	.118	.881
COMP	.187	.688	DMA	.380	.714	DMA	.424	.810
CE	.647	.675	COMP	.170	.601	CE	.657	.572
PS	.020	.609	CE	.546	.567	PS	-.115	.536
RDG	.054	.598	RDG	-.030	.516	GMRT	.159	.415

Note. LDF = standardized LDF coefficient (relative weight in the linear discriminant function standardized to equate unit of measurement; Loading = discriminant loading (correlation between the predictor variable and the standardized coefficients.

Classification

To classify new students, the LCF coefficients would be used linearly. In a two-group PDA, two LCFs were produced. The two functions generated for the three analyses completed were:

Analysis One:

$$LCF_{1a} = -70.121 + 1.231(GPA) + .130(COMP) + .151(CE) + .003(PS) + .253(RDG)$$

$$LCF_{1b} = -91.227 + 2.228(GPA) + .143\ (COMP) + .199(CE) + .004(PS) + .257(RDG)$$

Analysis Two:

$$LCF_{2a} = -91.064 + 2.746(GPA) - .483(DMA) + .141(COMP) + .248(CE) + .323(RDG)$$

$$LCF_{2b} = -112.374 + 3.693(GPA) - .379(DMA) + .154(COMP) + .296(CE) + .321(RDG)$$

Analysis Three:

$$LCF_{3a} = -303.618 + 4.210(GPA) - 1.423(DMA) + .331(CE) - .180(PS) + .1.034(GMRT)$$

$$LCF_{3a} = -329.565 + 4.496(GPA) - 1.294(DMA) + .392(CE) - .189(PS) + .1.047(GMRT)$$

To classify a new student, values for each predictor variable would be entered into both equations. The group on which the score is the highest would be the group in which the student is placed.

While unstandardized function coefficients have no direct interpretive value because they represent different metrics (Klecka, 1980), they can be multiplied by the associated values of the predictor variables and summed to form a set of discriminant scores. In a two-group solution these scores can then be graphed as a means of examining group separation, group dispersion relative to the centroids, and as a way of comparing how new cases fit within the context of existing groups. As an example, the histogram and related equation generated for Analysis One is presented in Figure 1.

This kind of representation provides a visual indication of group separation. As group separation increases, so does classification accuracy. The differences between the group centroids were 1.86, 2.02,

and 2.03 for Analyses One, Two, and Three, respectively indicating that the three analyses were comparable in terms of group separation. Additionally, the overlap between the Proficient and Not Proficient categories represents those scores that have a greater likelihood of misclassification. The graphs can be used as an indicator of both false negative and false positive misclassifications. Indeed, being able to differentiate between those who are proficient but are classified as not, and those who are not but are classified as proficient is of great value instructionally.

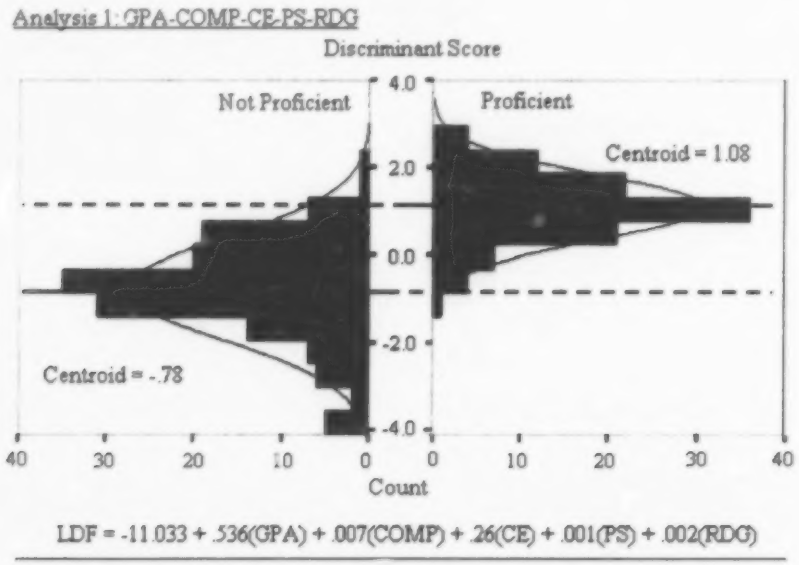

Figure 1. Horizontal histogram of discriminant scores by CSAP group.

Discussion

The initial question posed in this study related to the combination of performance indicators that best predicted scores on the CSAP. Of the three models presented above, Analysis Three (GPA-DMA-CE-PS-GMRT) was considered the best combination of the predictors used. Even with the smallest *N*, it generated the highest composite of accounted variance, classification accuracy, and mean posterior probability. Of more importance conceptually was the indicator variety seen within this analysis. This combination considered classroom performance (GPA) along with achievement on three different achievement measures for a total of four indicator types. Given that the conceptual strength of a set of multiple measures is drawn in large part from the types of indicators used, this combination was considered the best.

The broader issue addressed by this study concerned the use of PDA as a method of combining multiple measures of performance. The results clearly indicated that using a combination of multiple

BEST COPY AVAILABLE

measures was superior to relying on any single measure. Based on the analytical flexibility, amount and types of information generated, and the external validation ability, PDA was considered a viable and useful method of combining multiple measures as the basis for making appropriate inferences and decisions.

The analysis produced a great deal of interpretive information. For instance, the relationship between the predictor variables could be explored. The classification accuracies could be viewed alongside the posterior and conditional (typicality) probabilities to obtain a clearer view of group membership. The LDF distribution made it possible to look for outliers whose scores were considered anomalous and fence-riders who shared characteristics of both groups and were therefore prone to misclassification. The LCFs could then be used to classify new cases while maintaining a perspective on the distributional nature of the original group. The statistical and graphical information generated by a PDA could provide educators with the analytic details needed to make sound decisions.

Finally, the L-O-O external cross-validation procedure added to the level of confidence placed in the results. Even though the obtained results were very consistent across analyses, the presence of a conservative estimate of external validity strengthened the support of this method. Additionally, new cases could be classified with the model and then compared to the original group to identify possible changes in the overall dynamics and predictive power of a given combination.

Comparison With Other Methods

Various methods of combining multiple measures were described earlier. Some based judgment on the satisfactory performance on one, some, or all of the measures in the assessment battery. Since these methods are based only on component performance, they cannot account for the differences in test difficulty and cut-scores, nor can they include the interactions among the tests in the final judgment. Others computed weights from internal measures of the component variables such as the standard error of measurement. These methods are based on the distributional properties of the component variables and do not take into account the covariance between them. Both groups provide only for internal analyses and therefore lack measures of decision accuracy needed in educational settings. The results of this study demonstrated that a PDA addresses these concerns satisfactorily. The weights assigned were based on the interactions among the predictor variables and not on the relative difficulty of any one measure. Additionally, the PDA went beyond the internal distributions and generated a measure of predictive accuracy. These advantages provide a compelling argument supporting a PDA over the other methods.

Instructional Benefits of Multiple Measures

Isolated, high-stakes assessments have had a severe impact on school curricula. Of specific concern has been the narrowing of topics and strategies to match those covered by external assessments (Koretz, Linn, Dunbar, & Shepard, 1991). This narrowing has not only limited the range of academic

skills presented, but has also brought into question the validity of the assessments themselves due to the degree of strategy-specific instruction (Heubert & Hauser, 1999). By combining a wide range of assessments that contribute to a global estimate of proficiency, educators are encouraged to design their instructional units in a manner that fosters overall proficiency. Assessment-specific strategies would be counterproductive in that item formats would be secondary to the constructs in question. In this study, the application of a multiple measures approach was not only supported from a psychometric standpoint, it was also more consistent with a cognitive perspective (cf. Piaget, 1970; Snow & Lohman, 1989) that focuses on construct building and generalization, rather than skill demonstration as a means of establishing subject area proficiency.

Cautions and Caveats

Using the premise that (a) multiple measures are preferable to single observations, and (b) discriminant analysis is an effective tool for combining measures of educational performance, several issues need investigation before its use in high-stakes educational settings can be recommended. First, demographic generalizability needs attention. A given combination of educational indicators should be examined to verify that different groups of students (e.g., age, gender, ethnicity, economic status, etc.) are not treated unfairly.

Further work is also needed with respect to the variables included as predictors. The impact of reading performance should be explored and clarified. While I acknowledge that reading achievement impacted the PDA, its presence as a significant covariate made inferences about mathematics performance difficult. Instructionally, it is imperative that students with poor reading ability be allowed to demonstrate their true level of mathematics proficiency. The inclusion of teacher judgments apart from classroom grades should also be considered. Specific issues such as student responsibility and work effort may be important predictors of general performance. Similarly, teacher evaluations related to the level of proficiency attained by students should be developed. The opinions of classroom teachers with respect to student performance are generally limited to grades and other parent communications. Including these assessments systematically may add yet another type of predictor to a combination of multiple measures that effectively classifies student performance.

Methodologically, the effects of the possible violation of PDA assumptions need to be explored in the context of educational decisions. Many variables in educational settings are distributionally non-normal. If reasonable transformations have little or no effect, the impact on the results needs to be known if high-stakes decisions are to be made. The impact of unequal covariance matrices also needs exploration. This study deferred to the log determinants of the covariance matrices over the significance of Box's M (see SPSS, 1998b); the appropriateness of this decision should be explored.

Also of interest is the stability of the model over time. The ability to classify new students is of fundamental importance to this method. However, educators must recognize when the LDFs no longer reflect the true dynamics of a student population. It is reasonable to expect that component weights and interactions will change over time due to differences in student populations and changes in the instructional program. Procedures need to be developed to recognize to such changes and respond appropriately.

Where Do We Go From Here?

Notwithstanding the cautions mentioned above, the practical applications of the results deserve consideration. This study investigated PDA as a specific method of combining multiple measures of performance wherein the grouping variable was itself a measure of academic performance. This arrangement facilitated the evaluation of the method and the information provided. Ideally, the CSAP would be included as a predictor variable and an overall evaluation of grade level mathematics proficiency would be the grouping variable. With respect to this study, I had inferential concerns about the CSAP given that the only validity evidence reported to date supported an a priori alignment between the items and the standards based on expert opinion (CTB/McGraw-Hill, 2001). Neither evidence confirming this alignment, nor any concurrent or construct validity studies have been reported to date. If research supporting the use of the CSAP as an accurate measure of performance relative to the Colorado Model Content Standards (CDE, 1995) is made available, its use as a predictor in a model that classifies mathematics proficiency has merit. However until evidence is presented that supports the CSAP as an accurate measure of the content standards it was designed to reflect, its use in a decision-making context is inappropriate.

More generally however is a sort of chicken-and-egg problem. If the intent is to use a set of multiple measures to predict proficiency, how do we categorize the initial set of proficient and not proficient cases? The answer to this question may be necessarily iterative. To start the process, educators could develop an agreed upon set of proficient and not proficient cases. Other cases could then be classified using the LCFs obtained. LDF distributions (cf. Figure 1) would be examined to identify cases that may have been misclassified. The PDA could then be rerun and the process repeated until classifications produced by the model are congruent with those made empirically. Once the model stabilizes, new cases can be classified and an ongoing process of model evaluation and accuracy assessment can be initiated.

Summary

The interest in public school accountability has grown tremendously in recent years. This interest has resulted in a marked increase in large-scale assessment systems intended to measure student progress in various content areas. While policymakers and others outside the field of education have promoted

individual, discrete assessments as the sole indicators of student performance, those within the field have promoted comprehensive assessment systems that evaluate performance by combining different types of measures over time. The concept of using multiple measures of student performance is supported psychometrically (AERA, APA, & NCME, 1999), as well as pedagogically wherein classroom teachers regularly evaluate student work in the context of instruction. The use of single-instrument, large-scale assessments alone is inferior to a multiple measures approach in terms of current measurement standards, as well as from the cognitive perspective that isolated assessments cannot improve learning if they are detached from instruction.

In spite of the current rhetoric calling for isolated, external measures of educational accountability that often lack sufficient use-specific validity, important decisions related to student performance should be based on relevant indicators that reflect performance. These indicators should be combined in a manner that satisfies the conditions of being objective, transparent, and informative with respect to classification accuracy. The results of this study supported the contention that a predictive discriminant analysis can be an effective method of combining educational indicators for the purposes of providing information and making important student performance decisions in a manner that is consistent with these conditions.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Buras, A. (1996, Jan.). *Descriptive versus predictive discriminant analysis: A comparison and contrast of the two techniques*. Paper presented at the Annual meeting of the Southwest Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 395 981)

Colorado Department of Education (1995). *Colorado Model Content Standards for Mathematics*. Denver, CO: Author.

Colorado Department of Education (2000). *Colorado Student Assessment Program, 8th - Grade Mathematics Test*. Denver, CO: Author.

CTB/McGraw-Hill (2001). *Colorado Student Assessment Program: Technical report 2000* [On-line]. Monterey, CA: Author. Available: www.cde.state.co.us/cdeassess/pubassess.htm

Dillon, W. R. & Goldstein, M. (1984). *Multivariate analysis: Methods and applications*. New York: John Wiley & Sons.

Education Week (1999, Jan. 11). *Quality Counts '99: Rewarding Results, Punishing Failure*.

Good, R. (2001). The discriminating power of a combination of multiple measures on a large-scale, standards-based mathematics assessment. *Dissertation Abstracts International, 62*(09). (University Microfilm No. 3025975)

Gulliksen, H. (1950). *Theory of Mental Tests*. New York: John Wiley & Sons.

Heubert, J.P. & Hauser, R.M. (1999). *High-stakes: testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Hirst, D. (1996). Error-rate estimation in multiple-group linear discriminant analysis. *Technometrics, 38*(4), 389-393.

Hohn, A. & Veitch, W. R. (1999, April). *The Colorado Basic Literacy Act: Multiple measures in action*. Paper presented at the Annual meeting of the American Educational Research Association. Montreal.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1993). *Iowa Test of Basic Skills, Forms K and L, Level 14*. Chicago, IL: Riverside Publishing Co.

Huberty, C. J. (1984). Issues in the use and interpretation of discriminant analysis. *Psychological Bulletin, 95*(1), 156-171.

Huberty, C. J. (1994). *Applied Discriminant Analysis*. New York: John Wiley & Sons.

Huberty, C. J. & Barton, R. M. (1989). An introduction to discriminant analysis. *Measurement and Evaluation in Counseling and Development,* 22, 158-168.

Jang, Y. (1998, April). *Implementing standards-based multiple measures for IASA, Title I accountability using TerraNova multiple assessment*. Paper presented at the Annual meeting of the American Educational Research Association, San Diego, CA.

Klecka, W. R. (1980). *Discriminant analysis* (Sage University Paper series on Quantitative Applications in the Social Sciences, No. 07-019). Beverly Hills, CA: Sage Publications.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. Presented in R. L. Linn (Chair), *Effects of High-Stakes Educational Testing on Instruction and Achievement*. Symposium presented at an annual meeting of the American Educational Research Association, Chicago, IL.

Lewis, A. (2000, April). *High-stakes testing: Trends and issues* (Policy Brief). Aurora, CO: Mid-continent Research for Education and Learning.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Glenview, IL: Scott, Foresman and Co.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie Reading Tests* (4th ed.). Iasca, IL: Riverside Publishing Co.

Mehrens, W. A. (1990). Combining evaluation data from multiple sources. In J. Millman and L. Darling-Hammond (Eds.), *The New Handbook of Teacher Evaluation* (pp. 322-334). Newbury Park, CA: Sage Publications.

Novak, J. R., Winters, L., & Flores, E. (2000, April). *Using multiple measures for accountability purposes: One district's experience*. Paper presented at an annual meeting of the American Educational Research Association, New Orleans, LA.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart, and Winston.

Piaget, J. (1970). *Genetic epistemology*. New York: W. W. Norton & Co.

Rudner, L. M. (2000). *Informed test component weighting*. Maryland Assessment Research Center for Education Success & ERIC Clearinghouse on Assessment and Evaluation.

Ryan, J. M., & Hess, R. K. (1999, April). *Issues, strategies, and procedures for combining data from multiple measures*. Paper presented at an annual meeting of the American Educational Research Association, Montreal.

Snow, R. E., & Lohman, D. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 263-331). New York: American Council on Education and Macmillan Publishing Co.

SPSS (1998a). *Statistical Package for the Social Sciences* (version 8.0). Chicago, IL: Author.

SPSS (1998b). *SPSS Base 8.0: Applications guide*. Chicago, IL: Author.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55*(4), 525-534.

Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103-118.

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40*(5), 663-705.

Whitaker, J. S. (1997, Jan.). Use of stepwise methodology in discriminant analysis. Paper presented at the Annual meeting of the Southwest Educational Research Association, Austin, TX.

Woldbeck, T. (1998, April). *Two types of discriminant analysis: NOT six of one, half a dozen of another*. Paper presented at the Annual meeting of the American Educational Research Association, San Diego, CA. (ERIC Document Reproduction Service No. ED 418 128)